JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2019, VOL. 28, NO. 2, 350–361
https://doi.org/10.1080/10618600.2018.1529601

Taylor & Francis
Taylor & Francis Group

MM Algorithms for Variance Components Models

Hua Zhou[a], Liuyi Hu[b], Jin Zhou[c], and Kenneth Lange[d]

[a]Department of Biostatistics, University of California, Los Angeles, CA; [b]Department of Statistics, North Carolina State University, Raleigh, NC; [c]Division of Epidemiology and Biostatistcis, University of Arizona, Tuscon, AZ; [d]Department of Human Genetics, University of California, Los Angeles, CA

ABSTRACT

Variance components estimation and mixed model analysis are central themes in statistics with applications in numerous scientific disciplines. Despite the best efforts of generations of statisticians and numerical analysts, maximum likelihood estimation (MLE) and restricted MLE of variance component models remain numerically challenging. Building on the minorization–maximization (MM) principle, this article presents a novel iterative algorithm for variance components estimation. Our MM algorithm is trivial to implement and competitive on large data problems. The algorithm readily extends to more complicated problems such as linear mixed models, multivariate response models possibly with missing data, maximum a posteriori estimation, and penalized estimation. We establish the global convergence of the MM algorithm to a Karush–Kuhn–Tucker point and demonstrate, both numerically and theoretically, that it converges faster than the classical EM algorithm when the number of variance components is greater than two and all covariance matrices are positive definite. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received August 2017
Revised June 2018

KEYWORDS
Global convergence; Linear mixed model (LMM); Matrix convexity; Maximum a posteriori (MAP) estimation; Minorization–maximization (MM); Multivariate response; Penalized estimation; Variance components model

1. Introduction

Variance components and linear mixed models (LMMs) are among the most potent tools in a statistician's toolbox, finding numerous applications in agriculture, biology, economics, genetics, epidemiology, and medicine. Given an observed $n \times 1$ response vector $\boldsymbol{y}$ and $n \times p$ predictor matrix $\boldsymbol{X}$, the simplest variance components model postulates that $\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega} = \sum_{i=1}^{m} \sigma_i^2 \boldsymbol{V}_i$, and the $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_m$ are m fixed positive semidefinite matrices. The parameters of the model can be divided into mean effects $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and variance components $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_m^2)$. Throughout we assume $\boldsymbol{\Omega}$ is positive definite. The extension to singular $\boldsymbol{\Omega}$ will not be pursued here. Estimation revolves around the log-likelihood function

$$L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) = -\frac{1}{2} \ln \det \boldsymbol{\Omega} - \frac{1}{2} (\boldsymbol{y} - \boldsymbol{X\beta})^{\mathrm{T}} \boldsymbol{\Omega}^{-1} (\boldsymbol{y} - \boldsymbol{X\beta}). \quad (1)$$

Among the commonly used methods for estimating variance components, maximum likelihood estimation (MLE) (Hartley and Rao 1967) and restricted (or residual) MLE (REML) (Harville 1977) are the most popular. REML first projects $\boldsymbol{y}$ to the null space of $\boldsymbol{X}$ and then estimates variance components based on the projected responses. If the columns of a matrix $\boldsymbol{B}$ span the null space of $\boldsymbol{X}^{\mathrm{T}}$, then REML estimates the σ_i^2 by maximizing the log-likelihood of the redefined response vector $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}$, which is normally distributed with mean $\boldsymbol{0}$ and covariance $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{B} = \sum_{i=1}^{m} \sigma_i^2 \boldsymbol{B}^{\mathrm{T}} \boldsymbol{V}_i \boldsymbol{B}$.

There exists a large literature on iterative algorithms for finding MLE and REML (Laird and Ware 1982; Lindstrom and Bates 1988, 1990; Harville and Callanan 1990; Callanan and Harville 1991; Bates and Pinheiro 1998; Schafer and Yucel 2002). Fitting variance components models remains a challenge in models with a large sample size n or a large number of variance components m. Newton's method (Lindstrom and Bates 1988) converges quickly but is numerically unstable owing to the non-concavity of the log-likelihood. Fisher's scoring algorithm replaces the observed information matrix in Newton's method by the expected information matrix and yields an ascent algorithm when safeguarded by step halving. However, the calculation and inversion of expected information matrices cost $O(mn^3) + O(m^3)$ flops and quickly become impractical for large n or m, unless $\boldsymbol{V}_i$ are low rank, block diagonal, or have other special structures. The expectation–maximization (EM) algorithm initiated by Dempster Laird, and Rubin (1977) is a third alternative (Laird and Ware 1982; Laird, Lange, and Stram 1987; Lindstrom and Bates 1988; Bates and Pinheiro 1998). Compared to Newton's method, the EM algorithm is easy to implement and numerically stable, but painfully slow to converge. In practice, a strategy of priming Newton's method by a few EM steps leverages the stability of EM and the faster convergence of second-order methods.

In this article we derive a novel minorization–maximization (MM) algorithm for finding the MLE and REML estimates of variance components. We prove global convergence of the MM algorithm to a Karush–Kuhn–Tucker (KKT) point and explain why MM generally converges faster than EM for models with more than two variance components. We also sketch extensions of the MM algorithm to the multivariate response model with possibly missing responses, the LMM, maximum a posteriori (MAP) estimation, and penalized estimation. The numerical efficiency of the MM algorithm is illustrated through simulated datasets and a genomic example with 200 variance components.

CONTACT Hua Zhou huazhou@ucla.edu Department of Biostatistics, University of California, Los Angeles, CA 90095-1772.
Supplementary materials for this article are available online. Please go to www.tanfonline.com/r/JCGS.

2. Preliminaries

2.1. Background on MM Algorithms

Throughout we reserve Greek letters for parameters and indicate the current iteration number by a superscript t. The MM principle for maximizing an objective function $f(\boldsymbol{\theta})$ involves minorizing the objective function $f(\boldsymbol{\theta})$ by a surrogate function $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ around the current iterate $\boldsymbol{\theta}^{(t)}$ of a search (Lange, Hunter, and Yang 2000). Minorization is defined by the two conditions

$$\begin{aligned} f(\boldsymbol{\theta}^{(t)}) &= g(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) \\ f(\boldsymbol{\theta}) &\geq g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}), \qquad \boldsymbol{\theta} \neq \boldsymbol{\theta}^{(t)}. \end{aligned} \tag{2}$$

In other words, the surface $\boldsymbol{\theta} \mapsto g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ lies below the surface $\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta})$ and is tangent to it at the point $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$. Construction of the minorizing function $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ constitutes the first M of the MM algorithm. The second M of the algorithm maximizes the surrogate $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ rather than $f(\boldsymbol{\theta})$. The point $\boldsymbol{\theta}^{(t+1)}$ maximizing $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ satisfies the ascent property $f(\boldsymbol{\theta}^{(t+1)}) \geq f(\boldsymbol{\theta}^{(t)})$. This fact follows from the inequalities

$$f(\boldsymbol{\theta}^{(t+1)}) \geq g(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) \geq g(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) = f(\boldsymbol{\theta}^{(t)}), \tag{3}$$

reflecting the definition of $\boldsymbol{\theta}^{(t+1)}$ and the tangency and domination conditions (2). The ascent property makes the MM algorithm remarkably stable. The validity of the descent property depends only on increasing $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$, not on maximizing $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$. With obvious changes, the MM algorithm also applies to minimization rather than to maximization. To minimize a function $f(\boldsymbol{\theta})$, we majorize it by a surrogate function $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ and minimize $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ to produce the next iterate $\boldsymbol{\theta}^{(t+1)}$. The acronym should not be confused with the maximization–maximization algorithm in the variational Bayes context (Jeon 2012).

The MM principle (De Leeuw 1994; Heiser 1995; Lange, Hunter, and Yang 2000; Kiers 2002; Hunter and Lange 2004) finds applications in multidimensional scaling (Borg and Groenen 2005), ranking of sports teams (Hunter 2004), variable selection (Hunter and Li 2005; Yen 2011), optimal experiment design (Yu 2010), multivariate statistics (Zhou and Lange 2010), geometric programming (Lange and Zhou 2014), survival models (Hunter and Lange 2002; Ding, Tian, and Yuen 2015), sparse covariance estimation (Bien and Tibshirani 2011), and many other areas (Lange 2016). The celebrated EM principle (Dempster Laird, and Rubin 1977) is a special case of the MM principle. The Q function produced in the E step of an EM algorithm minorizes the log-likelihood up to an irrelevant constant. Thus, both EM and MM share the same advantages: simplicity, stability, graceful adaptation to constraints, and the tendency to avoid large matrix inversion. The more general MM perspective frees algorithm derivation from the missing data straitjacket and invites wider applications (Wu and Lange 2010). Figure 1 shows the minorization functions of EM and MM for a variance components model with $m = 2$ variance components.

2.2. Convex Matrix Functions

For symmetric matrices, we write $\boldsymbol{A} \preceq \boldsymbol{B}$ when $\boldsymbol{B} - \boldsymbol{A}$ is positive semidefinite and $\boldsymbol{A} \prec \boldsymbol{B}$ if $\boldsymbol{B} - \boldsymbol{A}$ is positive definite. A matrix-valued function f is said to be (matrix) convex if

$$f[\lambda \boldsymbol{A} + (1-\lambda)\boldsymbol{B}] \preceq \lambda f(\boldsymbol{A}) + (1-\lambda) f(\boldsymbol{B})$$

for all $\boldsymbol{A}$, $\boldsymbol{B}$, and $\lambda \in [0, 1]$. Our derivation of the MM variance components algorithm hinges on the convexity of the two functions mentioned in the next lemma. See standard text Boyd and Vandenberghe (2004) for the verification of both facts.

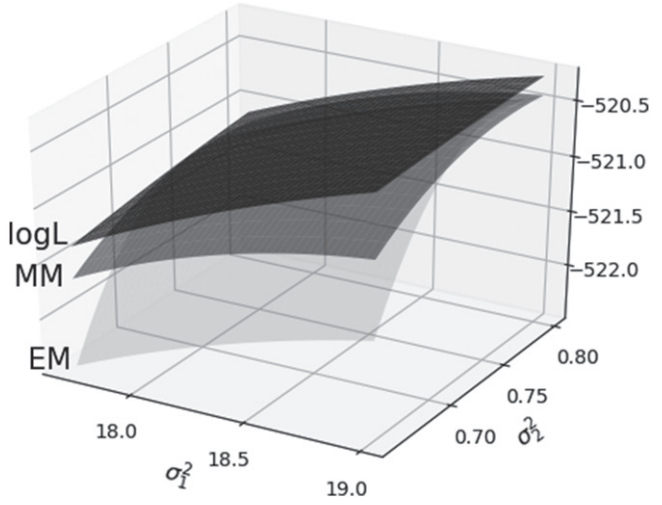

Figure 1. Surrogate functions of EM and MM minorize the log-likelihood surface of a 2-variance component model at point $(\sigma_1^{2(t)}, \sigma_2^{2(t)}) = (18.5, 0.7)$. MM surrogate function hugs the log-likelihood surface tighter than EM.

Lemma 1. (a) The matrix fractional function $f(\boldsymbol{A}, \boldsymbol{B}) = \boldsymbol{A}^{\mathrm{T}} \boldsymbol{B}^{-1} \boldsymbol{A}$ is jointly convex in the $m \times n$ matrix $\boldsymbol{A}$ and the $m \times m$ positive-definite matrix $\boldsymbol{B}$. (b) The log determinant function $f(\boldsymbol{B}) = \ln \det \boldsymbol{B}$ is concave on the set of positive-definite matrices.

3. Univariate Response Model

Our strategy for maximizing the log-likelihood (1) is to alternate updating the mean parameters $\boldsymbol{\beta}$ and the variance components $\boldsymbol{\sigma}^2$. Updating $\boldsymbol{\beta}$ given $\boldsymbol{\sigma}^2$ is a standard general least-squares problem with solution

$$\boldsymbol{\beta}^{(t+1)} = (\boldsymbol{X}^{\mathrm{T}} \boldsymbol{\Omega}^{-(t)} \boldsymbol{X})^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{\Omega}^{-(t)} \boldsymbol{y}, \tag{4}$$

where $\boldsymbol{\Omega}^{-(t)}$ represents the inverse of $\boldsymbol{\Omega}^{(t)} = \sum_{i=1}^m \sigma_i^{2(t)} \boldsymbol{V}_i$. Updating $\boldsymbol{\sigma}^2$ given $\boldsymbol{\beta}^{(t)}$ depends on two minorizations. If we assume that all of the $\boldsymbol{V}_i$ are positive definite, then the joint convexity of the map $(\boldsymbol{X}, \boldsymbol{Y}) \mapsto \boldsymbol{X}^{\mathrm{T}} \boldsymbol{Y}^{-1} \boldsymbol{X}$ for positive definite $\boldsymbol{Y}$ implies that

$$\begin{aligned} \boldsymbol{\Omega}^{(t)} \boldsymbol{\Omega}^{-1} \boldsymbol{\Omega}^{(t)} &= \left(\sum_{i=1}^m \sigma_i^{2(t)} \boldsymbol{V}_i \right) \left(\sum_{i=1}^m \sigma_i^2 \boldsymbol{V}_i \right)^{-1} \left(\sum_{i=1}^m \sigma_i^{2(t)} \boldsymbol{V}_i \right) \\ &\preceq \sum_{i=1}^m \frac{\sigma_i^{2(t)}}{\sum_j \sigma_j^{2(t)}} \left(\frac{\sum_j \sigma_j^{2(t)}}{\sigma_i^{2(t)}} \sigma_i^{2(t)} \boldsymbol{V}_i \right) \\ &\quad \times \left(\frac{\sum_j \sigma_j^{2(t)}}{\sigma_i^{2(t)}} \sigma_i^2 \boldsymbol{V}_i \right)^{-1} \left(\frac{\sum_j \sigma_j^{2(t)}}{\sigma_i^{2(t)}} \sigma_i^{2(t)} \boldsymbol{V}_i \right) \\ &= \sum_{i=1}^m \frac{\sigma_i^{4(t)}}{\sigma_i^2} \boldsymbol{V}_i. \end{aligned}$$

When one or more of the $\boldsymbol{V}_i$ are rank deficient, we replace each $\boldsymbol{V}_i$ by $\boldsymbol{V}_{i,\epsilon} = \boldsymbol{V}_i + \epsilon \boldsymbol{I}$ for $\epsilon > 0$ small and let $\boldsymbol{\Omega}_\epsilon^{(t)} = \sum_i \sigma_i^{2(t)} \boldsymbol{V}_{i,\epsilon}$. Sending ϵ to 0 in $\boldsymbol{\Omega}_\epsilon^{(t)} \boldsymbol{\Omega}_\epsilon^{-1} \boldsymbol{\Omega}_\epsilon^{(t)} \preceq \sum_{i=1}^m (\sigma_i^{4(t)}/\sigma_i^2) \boldsymbol{V}_{i,\epsilon}$ now gives the desired majorization $\boldsymbol{\Omega}^{(t)} \boldsymbol{\Omega}^{-1} \boldsymbol{\Omega}^{(t)} \preceq \sum_{i=1}^m (\sigma_i^{4(t)}/\sigma_i^2) \boldsymbol{V}_i$ in the general case. Negating both sides leads to the minorization

$$-(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{\Omega}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \geq -(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{\Omega}^{-(t)} \times \left(\sum_{i=1}^m \frac{\sigma_i^{4(t)}}{\sigma_i^2} \boldsymbol{V}_i \right) \boldsymbol{\Omega}^{-(t)} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{5}$$

that effectively separates the variance components $\sigma_1^2, \ldots, \sigma_m^2$ in the quadratic term of the log-likelihood (1).

The convexity of the function $\boldsymbol{A} \mapsto -\log\det \boldsymbol{A}$ is equivalent to the supporting hyperplane minorization

$$-\ln\det \boldsymbol{\Omega} \geq -\ln\det \boldsymbol{\Omega}^{(t)} - \mathrm{tr}[\boldsymbol{\Omega}^{-(t)} (\boldsymbol{\Omega} - \boldsymbol{\Omega}^{(t)})] \tag{6}$$

that separates $\sigma_1^2, \ldots, \sigma_m^2$ in the log determinant term of the log-likelihood (1). Combination of the minorizations (5) and (6) gives the overall minorization

$$\begin{aligned} & g(\boldsymbol{\sigma}^2 \mid \boldsymbol{\sigma}^{2(t)}) \\ &= -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Omega}^{-(t)} \boldsymbol{\Omega}) - \frac{1}{2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)})^{\mathrm{T}} \boldsymbol{\Omega}^{-(t)} \\ &\quad \times \left(\sum_{i=1}^m \frac{\sigma_i^{4(t)}}{\sigma_i^2} \boldsymbol{V}_i \right) \boldsymbol{\Omega}^{-(t)} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}) + c^{(t)} \qquad (7) \\ &= \sum_{i=1}^m \left[-\frac{\sigma_i^2}{2} \mathrm{tr}(\boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_i) - \frac{1}{2} \frac{\sigma_i^{4(t)}}{\sigma_i^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)})^{\mathrm{T}} \boldsymbol{\Omega}^{-(t)} \right. \\ &\qquad \left. \boldsymbol{V}_i \boldsymbol{\Omega}^{-(t)} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}) \right] + c^{(t)}, \end{aligned}$$

where $c^{(t)}$ is an irrelevant constant. Maximization of $g(\boldsymbol{\sigma}^2 \mid \boldsymbol{\sigma}^{2(t)})$ with respect to σ_i^2 yields the simple multiplicative update

$$\sigma_i^{2(t+1)} = \sigma_i^{2(t)} \sqrt{\frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)})^{\mathrm{T}} \boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_i \boldsymbol{\Omega}^{-(t)} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)})}{\mathrm{tr}(\boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_i)}}, \quad i = 1, \ldots, m. \tag{8}$$

As a sanity check on our derivation, consider the partial derivative

$$\frac{\partial}{\partial \sigma_i^2} L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) = -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Omega}^{-1} \boldsymbol{V}_i) + \frac{1}{2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{\Omega}^{-1} \boldsymbol{V}_i \boldsymbol{\Omega}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{9}$$

Given $\sigma_i^{2(t)} > 0$, it is clear from the update formula (8) that $\sigma_i^{2(t+1)} < \sigma_i^{2(t)}$ when $\frac{\partial}{\partial \sigma_i^2} L < 0$. Conversely $\sigma_i^{2(t+1)} > \sigma_i^{2(t)}$ when $\frac{\partial}{\partial \sigma_i^2} L > 0$.

Algorithm 1 summarizes the MM algorithm for MLE of the univariate response model (1). The update formula (8) assumes that the numerator under the square root sign is nonnegative and the denominator is positive. The numerator requirement is a consequence of the positive semidefiniteness of $\boldsymbol{V}_i$. The denominator requirement is not obvious but can be verified through the Hadamard (elementwise) product representation $\mathrm{tr}(\boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_i) = \mathbf{1}^{\mathrm{T}} (\boldsymbol{\Omega}^{-(t)} \odot \boldsymbol{V}_i) \mathbf{1}$. The following lemma of Schur (1911) is crucial. We give a self-contained probabilistic proof in supplementary materials S.1.

Input : $\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_m$
Output: MLE $\hat{\boldsymbol{\beta}}, \hat{\sigma}_1^2, \ldots, \hat{\sigma}_m^2$
Initialize $\sigma_i^{(0)} > 0, i = 1, \ldots, m$
repeat
  $\boldsymbol{\Omega}^{(t)} \leftarrow \sum_{i=1}^m \sigma_i^{2(t)} \boldsymbol{V}_i$
  $\boldsymbol{\beta}^{(t)} \leftarrow \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{\Omega}^{-(t)} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$
  $\boldsymbol{r}^{(t)} \leftarrow \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}$
  $\sigma_i^{2(t+1)} \leftarrow \sigma_i^{2(t)} \sqrt{(\boldsymbol{r}^{(t)\mathrm{T}} \boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_i \boldsymbol{\Omega}^{-(t)} \boldsymbol{r}^{(t)}) / \mathrm{tr}(\boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_i)}$, $i = 1, \ldots, m$
until *objective value converges*

Algorithm 1: MM algorithm for MLE of the variance components of model (1).

Lemma 2 (Schur). The Hadamard product of a positive-definite matrix with a positive semidefinite matrix with positive diagonal entries is positive definite.

We can now obtain the following characterization of the MM iterates.

Proposition 1. Assume $\boldsymbol{V}_i$ has strictly positive diagonal entries. Then $\mathrm{tr}(\boldsymbol{\Omega}^{-(t)} \boldsymbol{V}_i) > 0$ for all t. Furthermore if $\sigma_i^{2(0)} > 0$ and $\boldsymbol{\Omega}^{-(t)} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}) \notin \mathrm{null}(\boldsymbol{V}_i)$ for all t, then $\sigma_i^{2(t)} > 0$ for all t. When $\boldsymbol{V}_i$ is positive definite, $\sigma_i^{2(t)} > 0$ holds if and only if $\boldsymbol{y} \neq \boldsymbol{X}\boldsymbol{\beta}^{(t)}$.

Proof. The first claim follows easily from Schur's lemma. The second claim follows by induction. The third claim follows from the observation that $\mathrm{null}(\boldsymbol{V}_i) = \{\mathbf{0}\}$. □

In most applications, $\boldsymbol{V}_m = \boldsymbol{I}$. Proposition 1 guarantees that if $\sigma_m^{2(0)} > 0$ and the residual vector $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}$ is nonzero, then $\sigma_m^{2(t)}$ remains positive and thus $\boldsymbol{\Omega}^{(t)}$ remains positive definite throughout all iterations. This fact does not prevent any of the sequences $\sigma_i^{2(t)}$ from converging to 0. In this sense, the MM algorithm acts like an interior point method, approaching the optimum from inside the feasible region.

3.1. Univariate Response: Two Variance Components

The major computational cost of Algorithm 1 is inversion of the covariance matrix $\boldsymbol{\Omega}^{(t)}$ at each iteration. The special case of $m = 2$ variance components deserves attention as repeated matrix inversion can be avoided by invoking the simultaneous congruence decomposition for two symmetric matrices, one of which is positive definite (Rao 1973; Horn and Johnson 1985). This decomposition is also called the generalized eigenvalue decomposition (Golub and Van Loan 1996; Boyd and Vandenberghe 2004). If one assumes $\boldsymbol{\Omega} = \sigma_1^2 \boldsymbol{V}_1 + \sigma_2^2 \boldsymbol{V}_2$ and lets $(\boldsymbol{V}_1, \boldsymbol{V}_2) \mapsto (\boldsymbol{D}, \boldsymbol{U})$ be the decomposition with $\boldsymbol{U}$ nonsingular,

Input : $\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{V}_1, \boldsymbol{V}_2$
Output: MLE $\hat{\boldsymbol{\beta}}, \hat{\sigma}_1^2, \hat{\sigma}_2^2$
Simultaneous congruence decomposition: $(\boldsymbol{D}, \boldsymbol{U}) \leftarrow (\boldsymbol{V}_1, \boldsymbol{V}_2)$
Transform data: $\tilde{\boldsymbol{y}} \leftarrow \boldsymbol{U}^{\mathrm{T}}\boldsymbol{y}, \tilde{\boldsymbol{X}} \leftarrow \boldsymbol{U}^{\mathrm{T}}\boldsymbol{X}$
Initialize $\sigma_1^{(0)}, \sigma_1^{(0)} > 0$
repeat
  $w_i^{(t)} \leftarrow (\sigma_1^{2(t)} d_i + \sigma_2^{2(t)})^{-1}, \quad i = 1, \ldots, n$
  $\boldsymbol{\beta}^{(t)} \leftarrow \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n w_i^{(t)} (\tilde{y}_i - \tilde{\boldsymbol{x}}_i^{\mathrm{T}} \boldsymbol{\beta})^2$
  $\boldsymbol{r}^{(t)} \leftarrow \tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}^{(t)}$
  $\sigma_1^{2(t+1)} \leftarrow \sigma_1^{2(t)} \sqrt{\frac{\boldsymbol{r}^{(t)T}(\sigma_1^{2(t)}\boldsymbol{D}+\sigma_2^{2(t)}\boldsymbol{I})^{-1}\boldsymbol{D}(\sigma_1^{2(t)}\boldsymbol{D}+\sigma_2^{2(t)}\boldsymbol{I})^{-1}\boldsymbol{r}^{(t)}}{\mathrm{tr}[(\sigma_1^{2(t)}\boldsymbol{D}+\sigma_2^{2(t)}\boldsymbol{I})^{-1}\boldsymbol{D}]}}$
  $\sigma_2^{2(t+1)} \leftarrow \sigma_2^{2(t)} \sqrt{\frac{\boldsymbol{r}^{(t)T}(\sigma_1^{2(t)}\boldsymbol{D}+\sigma_2^{2(t)}\boldsymbol{I})^{-2}\boldsymbol{r}^{(t)}}{\mathrm{tr}[(\sigma_1^{2(t)}\boldsymbol{D}+\sigma_2^{2(t)}\boldsymbol{I})^{-1}]}}$
until *objective value converges*

Algorithm 2: Simplified MM algorithm for MLE of model (1) with $m = 2$ variance components and $\boldsymbol{\Omega} = \sigma_1^2\boldsymbol{V}_1 + \sigma_2^2\boldsymbol{V}_2$.

$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}_1\boldsymbol{U} = \boldsymbol{D}$ diagonal, and $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}_2\boldsymbol{U} = \boldsymbol{I}$, then

$$\begin{aligned} \boldsymbol{\Omega}^{(t)} &= \boldsymbol{U}^{-T}(\sigma_1^{2(t)}\boldsymbol{D} + \sigma_2^{2(t)}\boldsymbol{I}_n)\boldsymbol{U}^{-1} \\ \boldsymbol{\Omega}^{-(t)} &= \boldsymbol{U}(\sigma_1^{2(t)}\boldsymbol{D} + \sigma_2^{2(t)}\boldsymbol{I}_n)^{-1}\boldsymbol{U}^{\mathrm{T}} \\ \det(\boldsymbol{\Omega}^{(t)}) &= \det(\sigma_1^{2(t)}\boldsymbol{D} + \sigma_2^{2(t)}\boldsymbol{I}_n)\det(\boldsymbol{V}_2). \end{aligned} \tag{10}$$

With the revised responses $\tilde{\boldsymbol{y}} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{y}$ and the revised predictor matrix $\tilde{\boldsymbol{X}} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{X}$, the update (8) requires only vector operations and costs $O(n)$ flops. Updating the fixed effects is a weighted least-squares problem with the transformed data $(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{X}})$ and observation weights $w_i^{(t)} = (\sigma_1^{2(t)}d_i + \sigma_2^{2(t)})^{-1}$. Algorithm 2 summarizes the simplified MM algorithm for two variance components.

3.2. Numerical Experiments

This section compares the numerical performance of MM, EM, Fisher scoring (FS), and the `lme4` package in R (Bates et al. 2015) on simulated data from a two-way ANOVA random effects model and a genetic model. For ease of comparison, all algorithm runs start from $\boldsymbol{\sigma}^{2(0)} = \mathbf{1}$ and terminate when the relative change $(L^{(t+1)} - L^{(t)})/(|L^{(t)}| + 1)$ in the log-likelihood is less than 10^{-8}.

Two-way ANOVA: We simulated data from a two-way ANOVA random effects model

$$\begin{aligned} y_{ijk} &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \\ &1 \le i \le a,\ 1 \le j \le b,\ 1 \le k \le c, \end{aligned}$$

where $\alpha_i \sim N(0, \sigma_1^2)$, $\beta_j \sim N(0, \sigma_2^2)$, $(\alpha\beta)_{ij} \sim N(0, \sigma_3^2)$, and $\epsilon_{ijk} \sim N(0, \sigma_e^2)$ are jointly independent. Here, i indexes levels in factor 1, j indexes levels in factor 2, and k indexes observations in the (i, j)-combination. This corresponds to $m = 4$ variance components. In the simulation, we set $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$ and varied the ratio σ_1^2/σ_e^2; the numbers of levels a and b in factor 1 and factor 2, respectively; and the number of observations c in each combination of factor levels. For each simulation scenario, we simulated 50 replicates. The sample size was $n = abc$ for each replicate.

Tables 1 and 2 show the average number of iterations and the average runtimes when there are $a = b = 5$ levels of each factor. Based on these results and further results not shown for other combinations of a and b, we draw the following conclusions: FS takes the fewest iterations; the MM algorithm always takes fewer iterations than the EM algorithm; the faster rate of convergence of FS is outweighed by the extra cost of evaluating and inverting the information matrix. Table 1 in supplementary materials S.2 shows that all algorithms converged to the same objective values.

Genetic model: We simulated a quantitative trait $\boldsymbol{y}$ from a genetic model with two variance components and covariance matrix $\boldsymbol{\Omega} = \sigma_a^2\widehat{\boldsymbol{\Phi}} + \sigma_e^2\boldsymbol{I}$, where $\widehat{\boldsymbol{\Phi}}$ is a full-rank empirical kinship matrix estimated from the genome-wide measurements of 212 individuals using Option 29 of the Mendel software (Lange et al. 2013). In this example, MM had run times similar to FS, and both were much faster than EM and `lme4`.

In summary, the MM algorithm appears competitive even in small-scale examples. Many applications involve a large number of variance components. In this setting, the EM algorithm suffers from slow convergence and FS from an extremely high cost per iteration. Our genomic example in Section 7 reinforces this point.

4. Global Convergence of the MM Algorithm

The KKT necessary conditions for a local maximum $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_m^2)$ of the log-likelihood (1) require each component of the score vector to satisfy

$$\frac{\partial}{\partial \sigma_i^2} L(\boldsymbol{\sigma}^2) \in \begin{cases} \{0\} & \sigma_i^2 > 0 \\ (-\infty, 0] & \sigma_i^2 = 0\,. \end{cases}$$

In this section, we establish the global convergence of Algorithm 1 to a KKT point. To reduce the notational burden, we assume that $\boldsymbol{X}$ is null and omit estimation of fixed effects $\boldsymbol{\beta}$. The analysis easily extends to the nontrivial $\boldsymbol{X}$ case. Our convergence analysis relies on characterizing the properties of the objective function $L(\boldsymbol{\sigma}^2)$ and the MM algorithmic mapping $\boldsymbol{\sigma}^2 \mapsto M(\boldsymbol{\sigma}^2)$ defined by Equation (8). Special attention must be paid to the boundary values $\sigma_i^2 = 0$. We prove convergences for two cases, which cover most applications. For example, the genetic model in Section 3 satisfies Assumption 1, while the two-way ANOVA model satisfies Assumption 2.

Assumption 1. All $\boldsymbol{V}_i$ are positive definite.

Assumption 2. $\boldsymbol{V}_1$ is positive definite, each $\boldsymbol{V}_i$ is nontrivial, $\mathcal{H} = \mathrm{span}\{\boldsymbol{V}_2, \ldots, \boldsymbol{V}_m\}$ has dimension $q < n$, and $\boldsymbol{y} \notin \mathcal{H}$.

The key condition $\boldsymbol{y} \notin \mathrm{span}\{\boldsymbol{V}_2, \ldots, \boldsymbol{V}_m\}$ in the second case is also necessary for the existence of an MLE or REML (Demidenko and Massam 1999; Grzadziel and Michalski 2014). In supplementary materials, we derive a sequence of lemmas en route to the global convergence result declared in Theorem 1.

Theorem 1. Under either Assumption 1 or 2, the MM sequence $\{\boldsymbol{\sigma}^{2(t)}\}_{t\ge 0}$ has at least one limit point. Every limit point is a fixed

Table 1. FS converges fastest and MM takes fewer iterations than EM.

σ_1^2/σ_e^2	Method	c = # observations per combination			
		5	10	20	50
0.00	MM	143.12(99.76)	118.26(62.91)	96.26(50.61)	81.10(33.42)
	EM	2297.72(797.95)	1711.70(485.92)	1170.06(365.48)	788.10(216.60)
	FS	25.64(11.20)	21.10(7.00)	16.46(4.37)	13.88(2.88)
0.05	MM	121.86(98.52)	69.38(50.23)	55.88(37.34)	29.50(18.80)
	EM	1464.26(954.27)	538.04(504.42)	254.90(253.86)	104.98(157.97)
	FS	16.78(9.13)	12.62(6.22)	9.68(3.22)	8.10(1.34)
0.10	MM	84.74(59.33)	62.98(50.48)	40.46(31.43)	25.86(18.79)
	EM	985.46(830.49)	360.32(462.62)	157.70(231.91)	68.26(107.85)
	FS	15.20(10.10)	10.58(5.92)	8.58(3.56)	7.50(1.72)
1.00	MM	31.04(33.27)	29.60(27.66)	25.32(25.39)	24.90(20.76)
	EM	130.18(299.03)	161.14(290.23)	64.20(135.38)	84.88(137.88)
	FS	6.62(4.72)	6.32(3.64)	5.12(1.87)	5.36(1.50)
10.00	MM	29.80(35.42)	34.16(38.25)	28.82(28.44)	20.90(14.28)
	EM	115.94(274.33)	177.30(301.71)	80.12(155.67)	75.02(127.38)
	FS	12.72(5.14)	12.86(4.94)	11.66(3.95)	11.76(3.66)
20.00	MM	30.10(32.92)	32.72(39.02)	23.70(21.20)	19.62(15.67)
	EM	148.04(318.40)	85.86(180.28)	61.74(140.84)	37.36(83.89)
	FS	18.76(7.51)	17.40(5.21)	17.22(5.67)	16.28(5.03)

NOTE: Shown above are average number of iterations until convergence for MM, EM, and FS for fitting a two-way ANOVA model with $a = b = 5$ levels of both factors. Standard errors are given in parentheses.

Table 2. MM shows shortest run times than EM, FS, and `lme4`.

σ_1^2/σ_e^2	Method	c = # observations per combination			
		5	10	20	50
0.00	MM	11.46(7.77)	10.06(5.29)	11.93(6.35)	10.44(3.99)
	EM	189.32(71.32)	148.20(48.13)	147.87(49.97)	96.28(24.97)
	FS	34.27(33.47)	24.89(8.55)	23.70(14.15)	20.46(4.54)
	lme4	25.84(12.10)	22.32(1.25)	27.34(4.06)	36.14(5.59)
0.05	MM	9.79(7.72)	6.19(4.22)	6.87(4.37)	4.45(2.20)
	EM	116.03(75.57)	47.72(45.35)	30.60(29.88)	14.23(19.68)
	FS	19.18(10.23)	15.37(7.48)	12.78(4.06)	12.39(2.35)
	lme4	22.76(1.96)	24.88(2.60)	28.72(3.10)	47.34(16.29)
0.10	MM	7.07(4.78)	6.29(4.94)	5.14(3.72)	3.95(2.23)
	EM	78.96(66.19)	35.48(45.81)	19.53(27.71)	9.67(13.56)
	FS	17.36(11.26)	14.44(9.00)	12.08(6.31)	11.47(2.40)
	lme4	22.66(1.83)	28.90(8.70)	30.16(4.43)	44.58(4.89)
1.00	MM	2.66(2.61)	3.22(2.91)	3.57(3.15)	3.85(2.50)
	EM	10.71(23.93)	15.88(27.52)	8.35(16.26)	11.34(16.65)
	FS	7.88(5.44)	9.10(4.95)	7.12(2.42)	8.46(2.27)
	lme4	23.12(1.75)	30.22(9.37)	29.96(4.47)	42.82(8.32)
10.00	MM	2.48(2.72)	3.24(3.19)	3.84(3.35)	3.35(1.71)
	EM	9.66(22.02)	15.98(26.57)	10.24(18.78)	10.27(15.40)
	FS	15.19(6.05)	16.39(6.11)	15.81(5.15)	18.14(5.46)
	lme4	35.02(3.83)	47.12(8.10)	63.24(15.33)	102.78(34.49)
20.00	MM	2.57(2.49)	3.13(3.53)	3.13(2.44)	3.07(1.81)
	EM	12.28(25.71)	8.44(16.89)	8.01(17.12)	5.47(9.76)
	FS	22.09(8.53)	22.03(6.14)	23.08(7.21)	23.99(7.38)
	lme4	37.34(12.91)	50.24(8.59)	63.62(17.39)	91.14(28.39)

NOTE: Shown above are average run times (milliseconds) for fitting a two-way ANOVA model with $a = b = 5$ levels of both factors. Standard errors are given in parentheses.

point of $M(\boldsymbol{\sigma}^2)$. If the set of fixed points is discrete, then the MM sequence converges to one of them. Finally, when the iterates converge, their limit is a KKT point.

5. MM Versus EM

Examination of Tables 2 and 3 suggests that the MM algorithm usually converges faster than the EM algorithm. We now provide an explanation for this observation. Again for notational convenience, we consider the REML case where $\boldsymbol{X}$ is null. Since the EM principle is just a special instance of the MM principle, we can compare their convergence properties in a unified framework. Consider an MM map $M(\boldsymbol{\theta})$ for maximizing the objective function $f(\boldsymbol{\theta})$ via the surrogate function $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$. Close to the optimal point $\boldsymbol{\theta}^{\infty}$,

$$\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{\infty} \approx dM(\boldsymbol{\theta}^{\infty})(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{\infty}),$$

Table 3. MM and FS show superior performance than EM and `lme4`.

σ_a^2/σ_e^2	Method	Iteration	Runtime (ms)	Objective
0.00	MM	198.02(102.23)	133.61(822.67)	−375.59(9.63)
	EM	1196.10(958.51)	29.71(12.34)	−375.60(9.64)
	FS	7.60(3.07)	19.34(33.77)	−375.59(9.63)
	lme4	–	401.02(142.04)	−375.59(9.64)
0.05	MM	185.86(99.41)	17.26(1.76)	−377.39(10.52)
	EM	1227.62(1030.07)	29.82(12.74)	−377.40(10.52)
	FS	7.84(2.74)	14.97(1.55)	−377.39(10.52)
	lme4	–	425.04(144.00)	−377.39(10.52)
0.10	MM	169.24(99.75)	16.97(1.59)	−378.40(11.44)
	EM	924.80(912.23)	26.06(11.26)	−378.41(11.45)
	FS	7.32(2.75)	15.06(1.38)	−378.40(11.44)
	lme4	–	435.14(128.87)	−378.40(11.44)
1.00	MM	58.96(23.69)	15.53(0.75)	−409.54(10.90)
	EM	105.10(79.65)	15.49(0.96)	−409.54(10.90)
	FS	5.80(1.05)	14.66(0.89)	−409.54(10.90)
	lme4	–	493.14(52.80)	−409.54(10.90)
10.00	MM	110.00(63.13)	16.22(1.12)	−532.48(8.77)
	EM	642.48(1470.38)	22.32(18.37)	−532.57(8.75)
	FS	14.98(5.21)	14.78(0.97)	−531.72(8.92)
	lme4	–	2897.12(15006.38)	−532.48(8.77)
20.00	MM	110.52(34.81)	16.07(0.91)	−590.87(7.15)
	EM	1014.22(1775.40)	27.03(22.33)	−590.89(7.15)
	FS	17.72(3.13)	14.79(0.93)	−588.46(7.27)
	lme4	–	5059.24(20692.67)	−590.79(7.15)

NOTE: Shown above are average performance for fitting a genetic model. Standard errors are given in parentheses.

where $dM(\boldsymbol{\theta}^\infty)$ is the differential of the mapping M at the optimal point $\boldsymbol{\theta}^\infty$ of $f(\boldsymbol{\theta})$. Hence, the local convergence rate of the sequence $\boldsymbol{\theta}^{(t+1)} = M(\boldsymbol{\theta}^{(t)})$ coincides with the spectral radius of $dM(\boldsymbol{\theta}^\infty)$. Familiar calculations (Lange 2010) demonstrate that

$$dM(\boldsymbol{\theta}^\infty) = \boldsymbol{I} - [d^2 g(\boldsymbol{\theta}^\infty \mid \boldsymbol{\theta}^\infty)]^{-1} d^2 f(\boldsymbol{\theta}^\infty).$$

In other words, the local convergence rate is determined by how well the surrogate surface $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^\infty)$ approximates the objective surface $f(\boldsymbol{\theta})$ near the optimal point $\boldsymbol{\theta}^\infty$. In the EM literature, $dM(\boldsymbol{\theta}^\infty)$ is called the *rate matrix* (Meng and Rubin 1991). Fast convergence occurs when the surrogate $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^\infty)$ hugs the objective $f(\boldsymbol{\theta})$ tightly around $\boldsymbol{\theta}^\infty$. Figure 1 shows a case where the MM surrogate locally dominates the EM surrogate. We demonstrate that this is no accident.

The Q-function in the EM algorithm

$$g_{\text{EM}}(\boldsymbol{\sigma}^2 \mid \boldsymbol{\sigma}^{2(t)}) = -\frac{1}{2}\sum_{i=1}^m \left[\text{rank}(\boldsymbol{V}_i) \cdot \ln \sigma_i^2 + \text{rank}(\boldsymbol{V}_i)\frac{\sigma_i^{2(t)}}{\sigma_i^2} - \frac{\sigma_i^{4(t)}}{\sigma_i^2}\text{tr}(\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i)\right] - \frac{1}{2}\sum_{i=1}^m \frac{\sigma_i^{4(t)}}{\sigma_i^2}\boldsymbol{y}^{\text{T}}\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i\boldsymbol{\Omega}^{-(t)}\boldsymbol{y}$$

minorizes the log-likelihood up to an irrelevant constant. Supplementary materials S.6 gives a detailed derivation for the more general multivariate response case. Both surrogates $g_{\text{EM}}(\boldsymbol{\sigma}^2 \mid \boldsymbol{\sigma}^{2(\infty)})$ and $g_{\text{MM}}(\boldsymbol{\sigma}^2 \mid \boldsymbol{\sigma}^{2(\infty)})$ are parameter separated. This implies that both second differentials $d^2 g_{\text{EM}}(\boldsymbol{\sigma}^{2(\infty)} \mid \boldsymbol{\sigma}^{2(\infty)})$ and $d^2 g_{\text{MM}}(\boldsymbol{\sigma}^{2(\infty)} \mid \boldsymbol{\sigma}^{2(\infty)})$ are diagonal. A small diagonal entry of either matrix indicates fast convergence of the corresponding variance component. Our next result shows that, under Assumption 1, *on average* the diagonal entries of $d^2 g_{\text{EM}}(\boldsymbol{\sigma}^{2(\infty)} \mid \boldsymbol{\sigma}^{2(\infty)})$ dominate those of $d^2 g_{\text{MM}}(\boldsymbol{\sigma}^{2(\infty)} \mid \boldsymbol{\sigma}^{2(\infty)})$ when $m > 2$. Thus, the EM algorithm tends to converge more slowly than the MM algorithm, and the difference is more pronounced as the number of variance components m grows. See supplementary materials S.4 for the proof.

Theorem 2. Let $\boldsymbol{\sigma}^{2(\infty)} \succ \boldsymbol{0}_m$ be a common limit point of the EM and MM algorithms. Then both second differentials $d^2 g_{\text{MM}}(\boldsymbol{\sigma}^{2(\infty)} \mid \boldsymbol{\sigma}^{2(\infty)})$ and $d^2 g_{\text{EM}}(\boldsymbol{\sigma}^{2(\infty)} \mid \boldsymbol{\sigma}^{2(\infty)})$ are diagonal with

$$d^2 g_{\text{EM}}(\boldsymbol{\sigma}^{2(\infty)} \mid \boldsymbol{\sigma}^{2(\infty)})_{ii} = -\frac{\text{rank}(\boldsymbol{V}_i)}{2\sigma_i^{4(\infty)}}$$

$$d^2 g_{\text{MM}}(\boldsymbol{\sigma}^{2(\infty)} \mid \boldsymbol{\sigma}^{2(\infty)})_{ii} = -\frac{\boldsymbol{y}^{\text{T}}\boldsymbol{\Omega}^{-(\infty)}\boldsymbol{V}_i\boldsymbol{\Omega}^{-(\infty)}\boldsymbol{y}}{\sigma_i^{2(\infty)}} = -\frac{\text{tr}(\boldsymbol{\Omega}^{-(\infty)}\boldsymbol{V}_i)}{\sigma_i^{2(\infty)}}.$$

Furthermore, the average ratio

$$\frac{1}{m}\sum_{i=1}^m \frac{d^2 g_{\text{MM}}(\boldsymbol{\sigma}^{2(\infty)} \mid \boldsymbol{\sigma}^{2(\infty)})_{ii}}{d^2 g_{\text{EM}}(\boldsymbol{\sigma}^{2(\infty)} \mid \boldsymbol{\sigma}^{2(\infty)})_{ii}} = \frac{2}{mn}\sum_{i=1}^m \text{tr}(\boldsymbol{\Omega}^{-(\infty)}\sigma_i^{2(\infty)}\boldsymbol{V}_i) = \frac{2}{m} < 1$$

for $m > 2$ when all $\boldsymbol{V}_i$ have full rank n.

It is not clear whether a similar result holds under Assumption 2. Empirically, we observed faster convergence of MM than EM, for example, in the two-way ANOVA example

(Table 1). Also note that both the EM and MM algorithms must evaluate the traces $\mathrm{tr}(\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i)$ and quadratic forms $(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})^{\mathrm{T}}\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i\boldsymbol{\Omega}^{-(t)}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})$ at each iteration. Since these quantities are also the building blocks of the approximate rate matrices $d^2g(\boldsymbol{\sigma}^{2(t)}\,|\,\boldsymbol{\sigma}^{2(t)})$, one can rationally choose either the EM or MM updates based on which has smaller diagonal entries measured by the ℓ_1, ℓ_2, or ℓ_∞ norms. At negligible extra cost, this produces a hybrid algorithm that retains the ascent property and enjoys the better of the two convergence rates under either Assumption 1 or 2.

6. Extensions

Besides its competitive numerical performance, Algorithm 1 is attractive for its simplicity and ease of generalization. In this section, we outline MM algorithms for multivariate response models possibly with missing data, LMMs, MAP estimation, and penalized estimation.

6.1. Multivariate Response Model

Consider the multivariate response model with an $n \times d$ response matrix $\boldsymbol{Y}$, which has no missing entries, mean $\mathrm{E}\,\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B}$, and covariance

$$\boldsymbol{\Omega} = \mathrm{cov}(\mathrm{vec}\boldsymbol{Y}) \quad = \quad \sum_{i=1}^{m} \boldsymbol{\Gamma}_i \otimes \boldsymbol{V}_i.$$

The $p \times d$ coefficient matrix $\boldsymbol{B}$ collects the fixed effects, the $\boldsymbol{\Gamma}_i$ are unknown $d \times d$ variance components, and the $\boldsymbol{V}_i$ are known $n \times n$ covariance matrices. If the vector $\mathrm{vec}\boldsymbol{Y}$ is normally distributed, then $\boldsymbol{Y}$ equals a sum of independent matrix normal distributions (Gupta and Nagar 1999). We now make this assumption and pursue estimation of $\boldsymbol{B}$ and the $\boldsymbol{\Gamma}_i$, which we collectively denote as $\boldsymbol{\Gamma}$. Under the normality assumption, Roth's Kronecker product identity $\mathrm{vec}(\boldsymbol{CDE}) = (\boldsymbol{E}^{\mathrm{T}} \otimes \boldsymbol{C})\mathrm{vec}(\boldsymbol{D})$ yields the log-likelihood

$$\begin{aligned} L(\boldsymbol{B},\boldsymbol{\Gamma}) &= -\frac{1}{2}\ln\det\boldsymbol{\Omega} - \frac{1}{2}\mathrm{vec}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})^{\mathrm{T}}\boldsymbol{\Omega}^{-1} \\ &\quad \mathrm{vec}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}) \\ &= -\frac{1}{2}\ln\det\boldsymbol{\Omega} - \frac{1}{2}[\mathrm{vec}\boldsymbol{Y} - (\boldsymbol{I}_d \otimes \boldsymbol{X})\mathrm{vec}\boldsymbol{B}]^{\mathrm{T}} \\ &\quad \boldsymbol{\Omega}^{-1}[\mathrm{vec}\boldsymbol{Y} - (\boldsymbol{I}_d \otimes \boldsymbol{X})\mathrm{vec}\boldsymbol{B}]. \end{aligned} \tag{11}$$

Updating $\boldsymbol{B}$ given $\boldsymbol{\Gamma}^{(t)}$ is accomplished by solving the general least-squares problem met earlier in the univariate case. Update of $\boldsymbol{\Gamma}_i$ given $\boldsymbol{B}^{(t)}$ is difficult due to the positive semidefiniteness constraint. Typical solutions involve reparameterization of the covariance matrix (Pinheiro and Bates 1996). The MM algorithm derived in this section gracefully accommodates the covariance constraints.

Updating $\boldsymbol{\Gamma}$ given $\boldsymbol{B}^{(t)}$ requires generalizing the minorization (5). In view of Lemma 1 and the identities $(\boldsymbol{A}\otimes\boldsymbol{B})(\boldsymbol{C}\otimes\boldsymbol{D}) = (\boldsymbol{AC})\otimes(\boldsymbol{BD})$ and $(\boldsymbol{A}\otimes\boldsymbol{B})^{-1} = \boldsymbol{A}^{-1}\otimes\boldsymbol{B}^{-1}$, we have

$$\begin{aligned} \boldsymbol{\Omega}^{(t)}\boldsymbol{\Omega}^{-1}\boldsymbol{\Omega}^{(t)} &= m\left[\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\Gamma}_i^{(t)}\otimes\boldsymbol{V}_i\right]\left[\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\Gamma}_i\otimes\boldsymbol{V}_i\right]^{-1} \\ &\quad \times\left[\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\Gamma}_i^{(t)}\otimes\boldsymbol{V}_i\right] \\ &\preceq m\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{\Gamma}_i^{(t)}\otimes\boldsymbol{V}_i)(\boldsymbol{\Gamma}_i\otimes\boldsymbol{V}_i)^{-1}(\boldsymbol{\Gamma}_i^{(t)}\otimes\boldsymbol{V}_i) \\ &= \sum_{i=1}^{m}(\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{\Gamma}_i^{-1}\boldsymbol{\Gamma}_i^{(t)})\otimes\boldsymbol{V}_i, \end{aligned}$$

or equivalently

$$\boldsymbol{\Omega}^{-1} \preceq \boldsymbol{\Omega}^{-(t)}\Big[\sum_{i=1}^{m}(\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{\Gamma}_i^{-1}\boldsymbol{\Gamma}_i^{(t)})\otimes\boldsymbol{V}_i\Big]\boldsymbol{\Omega}^{-(t)}. \tag{12}$$

This derivation relies on the invertibility of the matrices $\boldsymbol{V}_i$. One can relax this assumption by substituting $\boldsymbol{V}_{\epsilon,i} = \boldsymbol{V}_i + \epsilon\boldsymbol{I}_n$ for $\boldsymbol{V}_i$ and sending ϵ to 0.

The majorization (12) and the minorization (6) jointly yield the surrogate

$$\begin{aligned} g(\boldsymbol{\Gamma}\mid\boldsymbol{\Gamma}^{(t)}) = -\frac{1}{2}\sum_{i=1}^{m}\Big\{&\mathrm{tr}[\boldsymbol{\Omega}^{-(t)}(\boldsymbol{\Gamma}_i\otimes\boldsymbol{V}_i)] \\ &+(\mathrm{vec}\,\boldsymbol{R}^{(t)})^{\mathrm{T}}[(\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{\Gamma}_i^{-1}\boldsymbol{\Gamma}_i^{(t)})\otimes\boldsymbol{V}_i]\,(\mathrm{vec}\,\boldsymbol{R}^{(t)})\Big\} \\ &+c^{(t)}, \end{aligned}$$

where $\boldsymbol{R}^{(t)}$ is the $n\times d$ matrix satisfying $\mathrm{vec}\,\boldsymbol{R}^{(t)} = \boldsymbol{\Omega}^{-(t)}\mathrm{vec}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}^{(t)})$ and $c^{(t)}$ is an irrelevant constant. Based on the Kronecker identities $(\mathrm{vec}\,\boldsymbol{A})^{\mathrm{T}}\mathrm{vec}\,\boldsymbol{B} = \mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{B})$ and $\mathrm{vec}(\boldsymbol{CDE}) = (\boldsymbol{E}^{\mathrm{T}}\otimes\boldsymbol{C})\mathrm{vec}(\boldsymbol{D})$, the surrogate can be rewritten as

$$\begin{aligned} g(\boldsymbol{\Gamma}\mid\boldsymbol{\Gamma}^{(t)}) &= -\frac{1}{2}\sum_{i=1}^{m}\Big\{\mathrm{tr}[\boldsymbol{\Omega}^{-(t)}(\boldsymbol{\Gamma}_i\otimes\boldsymbol{V}_i)] \\ &\qquad +\mathrm{tr}(\boldsymbol{R}^{(t)T}\boldsymbol{V}_i\boldsymbol{R}^{(t)}\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{\Gamma}_i^{-1}\boldsymbol{\Gamma}_i^{(t)})\Big\} + c^{(t)} \\ &= -\frac{1}{2}\sum_{i=1}^{m}\Big\{\mathrm{tr}[\boldsymbol{\Omega}^{-(t)}(\boldsymbol{\Gamma}_i\otimes\boldsymbol{V}_i)] \\ &\qquad +\mathrm{tr}(\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{R}^{(t)T}\boldsymbol{V}_i\boldsymbol{R}^{(t)}\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{\Gamma}_i^{-1})\Big\} + c^{(t)}. \end{aligned}$$

The first trace is linear in $\boldsymbol{\Gamma}_i$ with the coefficient of entry $(\boldsymbol{\Gamma}_i)_{jk}$ equal to

$$\mathrm{tr}(\boldsymbol{\Omega}_{jk}^{-(t)}\boldsymbol{V}_i) = \mathbf{1}_n^{\mathrm{T}}(\boldsymbol{V}_i\odot\boldsymbol{\Omega}_{jk}^{-(t)})\mathbf{1}_n,$$

where $\boldsymbol{\Omega}_{jk}^{-(t)}$ is the (j,k)th $n\times n$ block of $\boldsymbol{\Omega}^{-(t)}$ and $\odot$ indicates element-wise product. The matrix $\boldsymbol{M}_i$ of these coefficients can be written as

$$\boldsymbol{M}_i = (\boldsymbol{I}_d\otimes\mathbf{1}_n)^{\mathrm{T}}[(\mathbf{1}_d\mathbf{1}_d^{\mathrm{T}}\otimes\boldsymbol{V}_i)\odot\boldsymbol{\Omega}^{-(t)}](\boldsymbol{I}_d\otimes\mathbf{1}_n).$$

The directional derivative of $g(\boldsymbol{\Gamma}\mid\boldsymbol{\Gamma}^{(t)})$ with respect to $\boldsymbol{\Gamma}_i$ in the direction $\boldsymbol{\Delta}_i$ is

$$\begin{aligned} &-\frac{1}{2}\mathrm{tr}(\boldsymbol{M}_i\boldsymbol{\Delta}_i) + \frac{1}{2}\mathrm{tr}(\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{R}^{(t)T}\boldsymbol{V}_i\boldsymbol{R}^{(t)}\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{\Gamma}_i^{-1}\boldsymbol{\Delta}_i\boldsymbol{\Gamma}_i^{-1}) \\ &= -\frac{1}{2}\mathrm{tr}(\boldsymbol{M}_i\boldsymbol{\Delta}_i) + \frac{1}{2}\mathrm{tr}(\boldsymbol{\Gamma}_i^{-1}\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{R}^{(t)T}\boldsymbol{V}_i\boldsymbol{R}^{(t)}\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{\Gamma}_i^{-1}\boldsymbol{\Delta}_i). \end{aligned}$$

Because all directional derivatives of $g(\boldsymbol{\Gamma}\mid\boldsymbol{\Gamma}^{(t)})$ vanish at a stationary point, the matrix equation

$$\boldsymbol{M}_i = \boldsymbol{\Gamma}_i^{-1}\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{R}^{(t)T}\boldsymbol{V}_i\boldsymbol{R}^{(t)}\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{\Gamma}_i^{-1} \tag{13}$$

holds. Fortunately, this equation admits an explicit solution. For positive scalers a and b, the solution to the equation $b = x^{-1}ax^{-1}$ is $x = \pm\sqrt{a/b}$. The matrix analog of this equation

input : $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_m$
output: MLE $\widehat{\boldsymbol{B}}, \widehat{\boldsymbol{\Gamma}}_1, \ldots, \widehat{\boldsymbol{\Gamma}}_m$
Initialize $\boldsymbol{\Gamma}_i^{(0)}$ positive definite, $i = 1, \ldots, m$
repeat
  $\boldsymbol{\Omega}^{(t)} \leftarrow \sum_{i=1}^m \boldsymbol{\Gamma}_i^{(t)} \otimes \boldsymbol{V}_i$
  $\boldsymbol{B}^{(t)} \leftarrow \arg\min_{\boldsymbol{B}} [\text{vec}\boldsymbol{Y} - (\boldsymbol{I}_d \otimes \boldsymbol{X})\text{vec}\boldsymbol{B}]^{\mathrm{T}} \boldsymbol{\Omega}^{-(t)} [\text{vec}\boldsymbol{Y} - (\boldsymbol{I}_d \otimes \boldsymbol{X})\text{vec}\boldsymbol{B}]$
  $\boldsymbol{R}^{(t)} \leftarrow \text{reshape}(\boldsymbol{\Omega}^{-(t)}\text{vec}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}^{(t)}), n, d)$
  for $i = 1, \ldots, m$ **do**
    Cholesky $\boldsymbol{L}_i^{(t)} \boldsymbol{L}_i^{(t)T} \leftarrow (\boldsymbol{I}_d \otimes \boldsymbol{1}_n)^{\mathrm{T}} [(\boldsymbol{1}_d \boldsymbol{1}_d^{\mathrm{T}} \otimes \boldsymbol{V}_i) \odot \boldsymbol{\Omega}^{-(t)}](\boldsymbol{I}_d \otimes \boldsymbol{1}_n)$
    $\boldsymbol{\Gamma}_i^{(t+1)} \leftarrow \boldsymbol{L}_i^{-(t)T} [\boldsymbol{L}_i^{(t)T} (\boldsymbol{\Gamma}_i^{(t)} \boldsymbol{R}^{(t)T} \boldsymbol{V}_i \boldsymbol{R}^{(t)} \boldsymbol{\Gamma}_i^{(t)}) \boldsymbol{L}_i^{(t)}]^{1/2} \boldsymbol{L}_i^{-(t)}$
  end
until *objective value converges*

Algorithm 3: The MM algorithm for MLE of the multivariate response model (11).

is the Riccati equation $\boldsymbol{B} = \boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X}^{-1}$, whose solution is summarized in the next lemma.

Lemma 3. Assume $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive definite and $\boldsymbol{L}$ is the Cholesky factor of $\boldsymbol{B}$. Then $\boldsymbol{Y} = \boldsymbol{L}^{-T}(\boldsymbol{L}^T\boldsymbol{A}\boldsymbol{L})^{1/2}\boldsymbol{L}^{-1}$ is the unique positive-definite solution to the matrix equation $\boldsymbol{B} = \boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X}^{-1}$.

The Cholesky factor $\boldsymbol{L}$ in Lemma 3 can be replaced by the symmetric square root of $\boldsymbol{B}$. The solution, which is unique, remains the same. The Cholesky decomposition is preferred for its cheaper computational cost and better numerical stability.

Algorithm 3 summarizes the MM algorithm for fitting the multi-response model (3). Each iteration invokes m Cholesky decompositions and symmetric square roots of $d \times d$ positive-definite matrices. Fortunately in most applications, d is a small number. The following result guarantees the non-singularity of the Cholesky factor throughout the iterations. See supplementary materials S.8 for the proof.

Proposition 2. Assume $\boldsymbol{V}_i$ has strictly positive diagonal entries. Then the symmetric matrix $\boldsymbol{M}_i = (\boldsymbol{I}_d \otimes \boldsymbol{1}_n)^{\mathrm{T}}[(\boldsymbol{1}_d\boldsymbol{1}_d^{\mathrm{T}} \otimes \boldsymbol{V}_i) \odot \boldsymbol{\Omega}^{-(t)}](\boldsymbol{I}_d \otimes \boldsymbol{1}_n)$ is positive definite for all t. Furthermore if $\boldsymbol{\Gamma}_i^{(0)} \succ \boldsymbol{0}$ and no column of $\boldsymbol{R}^{(t)}$ lies in the null space of $\boldsymbol{V}_i$ for all t, then $\boldsymbol{\Gamma}_i^{(t)} \succ \boldsymbol{0}$ for all t.

6.2. Multivariate Response, Two Variance Components

When there are $m = 2$ variance components $\boldsymbol{\Omega} = \boldsymbol{\Gamma}_1 \otimes \boldsymbol{V}_1 + \boldsymbol{\Gamma}_2 \otimes \boldsymbol{V}_2$, repeated inversion of the $nd \times nd$ covariance matrix $\boldsymbol{\Omega}$ reduces to a single $n \times n$ simultaneous congruence decomposition and, per iteration, two $d \times d$ Cholesky decompositions and one $d \times d$ simultaneous congruence decomposition. The simultaneous congruence decomposition of the matrix pair $(\boldsymbol{V}_1, \boldsymbol{V}_2)$ involves generalized eigenvalues $\boldsymbol{d} = (d_1, \ldots, d_n)$ and a nonsingular matrix $\boldsymbol{U}$ such that $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}_1\boldsymbol{U} = \boldsymbol{D} = \text{diag}(\boldsymbol{d})$ and $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}_2\boldsymbol{U} = \boldsymbol{I}$. If the simultaneous congruence decomposition of $(\boldsymbol{\Gamma}_1^{(t)}, \boldsymbol{\Gamma}_2^{(t)})$ is $(\boldsymbol{\Lambda}^{(t)}, \boldsymbol{\Phi}^{(t)})$ with $\boldsymbol{\Phi}^{(t)T}\boldsymbol{\Gamma}_1^{(t)}\boldsymbol{\Phi}^{(t)} = \boldsymbol{\Lambda}^{(t)} = \text{diag}(\boldsymbol{\lambda}^{(t)})$ and $\boldsymbol{\Phi}^{(t)T}\boldsymbol{\Gamma}_2^{(t)}\boldsymbol{\Phi}^{(t)} = \boldsymbol{I}_d$, then

$$
\begin{aligned}
\boldsymbol{\Omega}^{(t)} &= (\boldsymbol{\Phi}^{-(t)} \otimes \boldsymbol{U}^{-1})^{\mathrm{T}}(\boldsymbol{\Lambda}^{(t)} \otimes \boldsymbol{D} + \boldsymbol{I}_d \otimes \boldsymbol{I}_n) \\
&\quad \times (\boldsymbol{\Phi}^{-(t)} \otimes \boldsymbol{U}^{-1}) \\
\boldsymbol{\Omega}^{-(t)} &= (\boldsymbol{\Phi}^{(t)} \otimes \boldsymbol{U})(\boldsymbol{\Lambda}^{(t)} \otimes \boldsymbol{D} + \boldsymbol{I}_d \otimes \boldsymbol{I}_n)^{-1}(\boldsymbol{\Phi}^{(t)} \otimes \boldsymbol{U})^{\mathrm{T}} \\
\det \boldsymbol{\Omega}^{(t)} &= \det(\boldsymbol{\Lambda}^{(t)} \otimes \boldsymbol{D} + \boldsymbol{I}_d \otimes \boldsymbol{I}_n) \\
&\quad \times \det[(\boldsymbol{\Phi}^{-(t)} \otimes \boldsymbol{U}^{-1})^{\mathrm{T}}(\boldsymbol{\Phi}^{-(t)} \otimes \boldsymbol{U}^{-1})] \\
&= \det(\boldsymbol{\Lambda}^{(t)} \otimes \boldsymbol{D} + \boldsymbol{I}_d \otimes \boldsymbol{I}_n) \det(\boldsymbol{\Gamma}_2^{(t)} \otimes \boldsymbol{V}_2) \\
&= \det(\boldsymbol{\Lambda}^{(t)} \otimes \boldsymbol{D} + \boldsymbol{I}_d \otimes \boldsymbol{I}_n) \det(\boldsymbol{\Gamma}_2^{(t)})^n \det(\boldsymbol{V}_2)^d.
\end{aligned}
$$

Updating the fixed effects reduces to a weighted least-squares problem for the transformed responses $\tilde{\boldsymbol{Y}} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{Y}$, transformed predictor matrix $\tilde{\boldsymbol{X}} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{X}$, and observation weights $(\lambda_k^{(t)} d_i + 1)^{-1}$. Algorithm 4 summarizes the simplified MM algorithm. The lengthy derivations are relegated to supplementary materials S.5.

Input : $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{V}_1, \boldsymbol{V}_2$
Output: MLE $\widehat{\boldsymbol{B}}, \widehat{\boldsymbol{\Gamma}}_1, \widehat{\boldsymbol{\Gamma}}_2$
Simultaneous congruence decomposition: $(\boldsymbol{D}, \boldsymbol{U}) \leftarrow (\boldsymbol{V}_1, \boldsymbol{V}_2)$
Transform data: $\tilde{\boldsymbol{Y}} \leftarrow \boldsymbol{U}^{\mathrm{T}}\boldsymbol{Y}, \tilde{\boldsymbol{X}} \leftarrow \boldsymbol{U}^{\mathrm{T}}\boldsymbol{X}$
Initialize $\boldsymbol{\Gamma}_1^{(0)}, \boldsymbol{\Gamma}_2^{(0)}$ positive definite
repeat
  Simultaneous congruence decomposition $(\boldsymbol{\Lambda}^{(t)}, \boldsymbol{\Phi}^{(t)}) \leftarrow (\boldsymbol{\Gamma}_1^{(t)}, \boldsymbol{\Gamma}_2^{(t)})$
  $\boldsymbol{B}^{(t)} \leftarrow \arg\min_{\boldsymbol{B}} [\text{vec}(\tilde{\boldsymbol{Y}}\boldsymbol{\Phi}^{(t)}) - (\boldsymbol{\Phi}^{(t)T} \otimes \tilde{\boldsymbol{X}})\text{vec}\boldsymbol{B}]^{\mathrm{T}}(\boldsymbol{\Lambda}^{(t)} \otimes \boldsymbol{D} + \boldsymbol{I}_d \otimes \boldsymbol{I}_n)^{-1}[\text{vec}(\tilde{\boldsymbol{Y}}\boldsymbol{\Phi}^{(t)}) - (\boldsymbol{\Phi}^{(t)T} \otimes \tilde{\boldsymbol{X}})\text{vec}\boldsymbol{B}]$
  Cholesky $\boldsymbol{L}_1^{(t)}\boldsymbol{L}_1^{(t)T} \leftarrow \boldsymbol{\Phi}^{(t)}\text{diag}\left(\text{tr}\left(\boldsymbol{D}(\lambda_k^{(t)}\boldsymbol{D} + \boldsymbol{I}_n)^{-1}\right), k = 1, \ldots, d\right)\boldsymbol{\Phi}^{(t)T}$
  Cholesky $\boldsymbol{L}_2^{(t)}\boldsymbol{L}_2^{(t)T} \leftarrow \boldsymbol{\Phi}^{(t)}\text{diag}\left(\text{tr}\left((\lambda_k^{(t)}\boldsymbol{D} + \boldsymbol{I}_n)^{-1}\right), k = 1, \ldots, d\right)\boldsymbol{\Phi}^{(t)T}$
  $\boldsymbol{N}_1^{(t)} \leftarrow \boldsymbol{D}^{1/2}[(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{B}^{(t)})\boldsymbol{\Phi}^{(t)}) \oslash (\boldsymbol{d}\boldsymbol{\lambda}^{(t)T} + \boldsymbol{1}_n\boldsymbol{1}_d^{\mathrm{T}})]\boldsymbol{\Lambda}^{(t)}\boldsymbol{\Phi}^{-(t)}$
  $\boldsymbol{N}_2^{(t)} \leftarrow [(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{B}^{(t)})\boldsymbol{\Phi}^{(t)}) \oslash (\boldsymbol{d}\boldsymbol{\lambda}^{(t)T} + \boldsymbol{1}_n\boldsymbol{1}_d^{\mathrm{T}})]\boldsymbol{\Phi}^{-(t)}$
  $\boldsymbol{\Gamma}_i^{(t+1)} \leftarrow \boldsymbol{L}_i^{-(t)T}(\boldsymbol{L}_i^{(t)T}\boldsymbol{M}_i^{(t)T}\boldsymbol{M}_i^{(t)}\boldsymbol{L}_i^{(t)})^{1/2}\boldsymbol{L}_i^{-(t)}, i = 1, 2$
until *objective value converges*

Algorithm 4: MM algorithm for multivariate response model $\boldsymbol{\Omega} = \boldsymbol{\Gamma}_1 \otimes \boldsymbol{V}_1 + \boldsymbol{\Gamma}_2 \otimes \boldsymbol{V}_2$ with two variance components matrices. Note that $\oslash$ denotes a Hadamard quotient.

6.3. Multivariate Response Model with Missing Responses

In many applications, the multivariate response model (11) involves missing responses. For instance, in testing multiple longitudinal traits in genetics, some trait values y_{ij} may be missing due to dropped patient visits, while their genetic covariates are complete. Missing data destroy the symmetry of the

log-likelihood (11) and complicates finding the MLE. Fortunately, MM Algorithm 3 easily adapts to this challenge.

The familiar EM argument (McLachlan and Krishnan 2008, sec. 2.2) shows that

$$-\frac{n}{2}\ln\det\boldsymbol{\Omega}^{(t)} - \frac{1}{2}\mathrm{tr}\{\boldsymbol{\Omega}^{-(t)}[\mathrm{vec}(\boldsymbol{Z}^{(t)}-\boldsymbol{X}\boldsymbol{B}^{(t)})\mathrm{vec}(\boldsymbol{Z}^{(t)}-\boldsymbol{X}\boldsymbol{B}^{(t)})^{\mathrm{T}}+\boldsymbol{C}^{(t)}]\} \tag{14}$$

minorizes the observed log-likelihood at the current iterate $(\boldsymbol{B}^{(t)}, \boldsymbol{\Gamma}_1^{(t)}, \ldots, \boldsymbol{\Gamma}_m^{(t)})$. Here, $\boldsymbol{Z}^{(t)}$ is the completed response matrix given the observed responses $\boldsymbol{Y}_{\text{obs}}^{(t)}$ and the current parameter values. The complete data $\boldsymbol{Y}$ is assumed to be normally distributed $N(\mathrm{vec}(\boldsymbol{X}\boldsymbol{B}^{(t)}), \boldsymbol{\Omega}^{(t)})$. The block matrix $\boldsymbol{C}^{(t)}$ is 0 except for a lower-right block consisting of a Schur complement.

To maximize the surrogate (14), we invoke the familiar minorization (6) and majorization (12) to separate the variance components $\boldsymbol{\Gamma}_i$. At each iteration, we impute missing entries by their conditional means, compute their conditional variances and covariances to supply the Schur complement, and then update the fixed effects and variance components by the explicit updates of Algorithm 3. The required conditional means and conditional variances can be conveniently obtained in the process of inverting $\boldsymbol{\Omega}^{(t)}$ by the sweep operator of computational statistics (Lange 2010, Section 7.3).

6.4. Linear Mixed Model

The LMM plays a central role in longitudinal data analysis. Consider the single-level LMM (Laird and Ware 1982; Bates and Pinheiro 1998) for n independent data clusters $(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ with

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{\beta}$ is a vector of fixed effects, the $\boldsymbol{\gamma}_i \sim N(\boldsymbol{0}, \boldsymbol{R}_i(\boldsymbol{\theta}))$ are random effects, and $\boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n_i})$ captures random noise independent of $\boldsymbol{\gamma}_i$. We assume the matrices $\boldsymbol{Z}_i$ have full column rank. The within-cluster covariance matrices $\boldsymbol{R}_i(\boldsymbol{\theta})$ depend on a parameter vector $\boldsymbol{\theta}$; typical choices for $\boldsymbol{R}_i(\boldsymbol{\theta})$ impose autocorrelation, compound symmetry, or unstructured correlation. It is clear that $\boldsymbol{Y}_i$ is normal with mean $\boldsymbol{X}_i\boldsymbol{\beta}$, covariance $\boldsymbol{\Omega}_i = \boldsymbol{Z}_i\boldsymbol{R}_i(\boldsymbol{\theta})\boldsymbol{Z}_i^{\mathrm{T}} + \sigma^2\boldsymbol{I}_{n_i}$, and log-likelihood

$$L_i(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = -\frac{1}{2}\ln\det\boldsymbol{\Omega}_i - \frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{\Omega}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}).$$

The next three technical facts about pseudo-inverses are used in deriving the MM algorithm for LMM and their proofs are in supplementary materials S.9–S.11.

Lemma 4. If $\boldsymbol{A}$ has full column rank and $\boldsymbol{B}$ has full row rank, then $(\boldsymbol{A}\boldsymbol{B})^+ = \boldsymbol{B}^+\boldsymbol{A}^+$.

Lemma 5. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive semidefinite matrices with the same range, then

$$\lim_{\epsilon\downarrow 0}(\boldsymbol{B}+\epsilon\boldsymbol{I})(\boldsymbol{A}+\epsilon\boldsymbol{I})^{-1}(\boldsymbol{B}+\epsilon\boldsymbol{I}) = \boldsymbol{B}\boldsymbol{A}^+\boldsymbol{B}.$$

Lemma 6. If $\boldsymbol{R}$ and $\boldsymbol{S}$ are positive-definite matrices, and the conformable matrix $\boldsymbol{Z}$ has full column rank, then the matrices $\boldsymbol{Z}\boldsymbol{R}\boldsymbol{Z}^{\mathrm{T}}$ and $\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}^{\mathrm{T}}$ share a common range.

The convexity of the map $(\boldsymbol{X}, \boldsymbol{Y}) \mapsto \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}^{-1}\boldsymbol{X}$ and Lemmas 4–6 now yield via the obvious limiting argument the majorization

$$\begin{aligned}\boldsymbol{\Omega}^{(t)}\boldsymbol{\Omega}^{-1}\boldsymbol{\Omega}^{(t)} &= (\boldsymbol{Z}_i\boldsymbol{R}_i(\boldsymbol{\theta}^{(t)})\boldsymbol{Z}_i^{\mathrm{T}} + \sigma^{2(t)}\boldsymbol{I}_{n_i})(\boldsymbol{Z}_i\boldsymbol{R}_i(\boldsymbol{\theta})\boldsymbol{Z}_i^{\mathrm{T}} + \sigma^2\boldsymbol{I}_{n_i})^{-1} \\ &\quad \times(\boldsymbol{Z}_i\boldsymbol{R}_i(\boldsymbol{\theta}^{(t)})\boldsymbol{Z}_i^{\mathrm{T}} + \sigma^{2(t)}\boldsymbol{I}_{n_i}) \\ &\preceq (\boldsymbol{Z}_i\boldsymbol{R}_i(\boldsymbol{\theta}^{(t)})\boldsymbol{Z}_i^{\mathrm{T}})(\boldsymbol{Z}_i\boldsymbol{R}_i(\boldsymbol{\theta})\boldsymbol{Z}_i^{\mathrm{T}})^{+}(\boldsymbol{Z}_i\boldsymbol{R}_i(\boldsymbol{\theta}^{(t)})\boldsymbol{Z}_i^{\mathrm{T}}) \\ &\quad +\frac{\sigma^{4(t)}}{\sigma^2}\boldsymbol{I}_{n_i} \\ &= [\boldsymbol{Z}_i\boldsymbol{R}_i(\boldsymbol{\theta}^{(t)})\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{Z}_i^{\mathrm{T}+}]\boldsymbol{R}_i^{-1}(\boldsymbol{\theta})[\boldsymbol{Z}_i^{+}\boldsymbol{Z}_i\boldsymbol{R}_i(\boldsymbol{\theta}^{(t)})\boldsymbol{Z}_i^{\mathrm{T}}] \\ &\quad +\frac{\sigma^{4(t)}}{\sigma^2}\boldsymbol{I}_{n_i}.\end{aligned}$$

In combination with the minorization (6), this gives the surrogate

$$\begin{aligned}&g_i(\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{\theta}^{(t)}, \sigma^{2(t)}) \\ &= -\frac{1}{2}\mathrm{tr}(\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\Omega}_i^{-(t)}\boldsymbol{Z}_i\boldsymbol{R}_i(\boldsymbol{\theta})) - \frac{1}{2}\boldsymbol{r}_i^{(t)T}\boldsymbol{R}_i^{-1}(\boldsymbol{\theta})\boldsymbol{r}_i^{(t)} \\ &\quad -\frac{\sigma^2}{2}\mathrm{tr}(\boldsymbol{\Omega}_i^{-(t)}) \\ &\quad -\frac{\sigma^{4(t)}}{2\sigma^2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}^{(t)})^{\mathrm{T}}\boldsymbol{\Omega}_i^{-2(t)}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}^{(t)}) + c^{(t)},\end{aligned}$$

for the log-likelihood $L_i(\boldsymbol{\theta}, \sigma^2)$, where

$$\begin{aligned}\boldsymbol{r}_i^{(t)} &= (\boldsymbol{Z}_i^{+}\boldsymbol{Z}_i\boldsymbol{R}_i(\boldsymbol{\theta}^{(t)})\boldsymbol{Z}_i^{\mathrm{T}})\boldsymbol{\Omega}_i^{-(t)}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}^{(t)}) \\ &= \boldsymbol{R}_i(\boldsymbol{\theta}^{(t)})\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\Omega}_i^{-(t)}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}^{(t)}).\end{aligned}$$

The parameters $\boldsymbol{\theta}$ and σ^2 are nicely separated. To maximize the overall minorization function $\sum_i g_i(\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{\theta}^{(t)}, \sigma^{2(t)})$, we update σ^2 via

$$\sigma^{2(t+1)} = \sigma^{2(t)}\sqrt{\frac{\sum_i(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}^{(t)})^{\mathrm{T}}\boldsymbol{\Omega}_i^{-2(t)}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}^{(t)})}{\sum_i \mathrm{tr}(\boldsymbol{\Omega}_i^{-(t)})}}.$$

For structured models such as autocorrelation and compound symmetry, updating $\boldsymbol{\theta}$ is a low-dimensional optimization problem that can be approached through the stationarity condition

$$\sum_i \mathrm{vec}\left(\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\Omega}_i^{(t)}\boldsymbol{Z}_i - \boldsymbol{R}_i^{-1}(\boldsymbol{\theta})\boldsymbol{r}_i^{(t)}\boldsymbol{r}_i^{(t)T}\boldsymbol{R}_i^{-1}(\boldsymbol{\theta})\right)^{\mathrm{T}}\frac{\partial}{\partial\theta_j}\mathrm{vec}\,\boldsymbol{R}_i(\boldsymbol{\theta}) = 0$$

for each component θ_j. For the unstructured model with $\boldsymbol{R}_i(\boldsymbol{\theta}) = \boldsymbol{R}$ for all i, the stationarity condition reads

$$\sum_i \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\Omega}_i^{(t)}\boldsymbol{Z}_i = \boldsymbol{R}^{-1}\left(\sum_i \boldsymbol{r}_i^{(t)}\boldsymbol{r}_i^{(t)T}\right)\boldsymbol{R}^{-1}$$

and admits an explicit solution based on Lemma 3.

The same tactics apply to a multilevel LMM (Bates and Pinheiro 1998) with responses

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_{i1}\boldsymbol{\gamma}_{i1} + \cdots \boldsymbol{Z}_{im}\boldsymbol{\gamma}_{im} + \boldsymbol{\epsilon}_i.$$

Minorization separates parameters for each level (variance component). Depending on the complexity of the covariance matrices, maximization of the surrogate can be accomplished analytically. For the sake of brevity, details are omitted.

6.5. MAP Estimation

Suppose $\boldsymbol{\beta}$ follows an improper flat prior, the variance components σ_i^2 follow inverse gamma priors with shapes $\alpha_i > 0$ and scales $\gamma_i > 0$, and these priors are independent. The log-posterior density then reduces to

$$-\frac{1}{2}\ln\det\boldsymbol{\Omega}-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) -\sum_{i=1}^{m}(\alpha_i+1)\ln\sigma_i^2-\sum_{i=1}^{m}\frac{\gamma_i}{\sigma_i^2}+c, \tag{15}$$

where c is an irrelevant constant. The MAP estimator of $(\boldsymbol{\beta},\boldsymbol{\sigma}^2)$ is the mode of the posterior distribution. The update (4) of $\boldsymbol{\beta}$ given $\boldsymbol{\sigma}^2$ remains the same. To update $\boldsymbol{\sigma}^2$ given $\boldsymbol{\beta}$, apply the same minorizations (5) and (6) to the first two terms of equation (15). This separates parameters and yields a convex surrogate for each σ_i^2. The minimum of the σ_i^2 surrogate is defined by the stationarity condition

$$\begin{aligned}0=&-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i)\\&+\frac{\sigma_i^{4(t)}}{2\sigma_i^4}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})^{\mathrm{T}}\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i\boldsymbol{\Omega}^{-(t)}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})\\&-\frac{\alpha_i+1}{\sigma_i^2}+\frac{\gamma_i}{\sigma_i^4}.\end{aligned}$$

Multiplying this by σ_i^4 gives a quadratic equation in σ_i^2. The positive root should be taken to meet the nonnegativity constraint on σ_i^2.

For the multivariate response model (11), we assume the variance components $\boldsymbol{\Gamma}_i$ follow independent inverse Wishart distributions with degrees of freedom $\nu_i > d-1$ and scale matrix $\boldsymbol{\Psi}_i \succ \boldsymbol{0}$. The log density of the posterior distribution is

$$-\frac{1}{2}\ln\det\boldsymbol{\Omega}-\frac{1}{2}\mathrm{vec}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathrm{vec}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}) -\frac{1}{2}\sum_{i=1}^{m}(\nu_i+d+1)\ln\det\boldsymbol{\Gamma}_i-\frac{1}{2}\sum_{i=1}^{m}\mathrm{tr}(\boldsymbol{\Psi}_i\boldsymbol{\Gamma}_i^{-1})+c, \tag{16}$$

where c is an irrelevant constant. Invoking the minorizations (6) and (12) for the first two terms and the supporting hyperplane minorization for $-\ln\det\boldsymbol{\Gamma}_i$ gives the surrogate function

$$\begin{aligned}g(\boldsymbol{\Gamma}|\boldsymbol{\Gamma}^{(t)})=&-\frac{1}{2}\sum_{i=1}^{m}\mathrm{tr}(\boldsymbol{\Omega}^{-(t)}(\boldsymbol{\Gamma}_i\otimes\boldsymbol{V}_i))\\&-\frac{1}{2}\sum_{i=1}^{m}\mathrm{tr}\left(\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{R}^{(t)T}\boldsymbol{V}_i\boldsymbol{R}^{(t)}\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{\Gamma}_i^{-1}\right)\\&-\frac{1}{2}\sum_{i=1}^{m}(\nu_i+d+1)\mathrm{tr}(\boldsymbol{\Gamma}_i^{-(t)}\boldsymbol{\Gamma}_i)\\&-\frac{1}{2}\sum_{i=1}^{m}\mathrm{tr}(\boldsymbol{\Psi}_i\boldsymbol{\Gamma}_i^{-1})+c^{(t)}.\end{aligned}$$

The optimal $\boldsymbol{\Gamma}_i$ satisfies the stationarity condition

$$\begin{aligned}&(\boldsymbol{I}_d\otimes\boldsymbol{1}_n)^{\mathrm{T}}[(\boldsymbol{1}_d\boldsymbol{1}_d^{\mathrm{T}}\otimes\boldsymbol{V}_i)\odot\boldsymbol{\Omega}^{-(t)}](\boldsymbol{I}_d\otimes\boldsymbol{1}_n)\\&\quad+(\nu_i+d+1)\boldsymbol{\Gamma}_i^{-(t)}\\&=\boldsymbol{\Gamma}_i^{-1}(\boldsymbol{\Gamma}_i^{(t)}\boldsymbol{R}^{(t)T}\boldsymbol{V}_i\boldsymbol{R}^{(t)}\boldsymbol{\Gamma}_i^{(t)}+\boldsymbol{\Psi}_i)\boldsymbol{\Gamma}_i^{-1},\end{aligned}$$

which can be solved by Lemma 3.

6.6. Variable Selection

In the statistical analysis of high-dimensional data, the imposition of sparsity leads to better interpretation and more stable parameter estimation. MM algorithms mesh well with penalized estimation. The simple variance components model (1) illustrates this fact. For the selection of fixed effects, minimizing the lasso-penalized log-likelihood $-L(\boldsymbol{\beta},\boldsymbol{\sigma}^2)+\lambda\sum_j|\beta_j|$ is often recommended (Schelldorfer, Bühlmann, and van de Geer 2011). The only change to the MM Algorithm 1 is that in estimating $\boldsymbol{\beta}$, one solves a lasso penalized general least-squares problem rather than an ordinary general least-squares problem. The updates of the variance components σ_i^2 remain the same. For estimation of a large number of variance components, one can minimize the ridge-penalized log-likelihood

$$-L(\boldsymbol{\beta},\boldsymbol{\sigma}^2)+\lambda\sum_{i=1}^{m}\sigma_i^2$$

subject to the nonnegativity constraints $\sigma_i^2 \geq 0$. The variance update (8) becomes

$$\sigma_i^{2(t+1)}=\sigma_i^{2(t)}\sqrt{\frac{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})^{\mathrm{T}}\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i\boldsymbol{\Omega}^{-(t)}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})}{\mathrm{tr}(\boldsymbol{\Omega}^{-(t)}\boldsymbol{V}_i)+2\lambda}},\quad i=1,\ldots,m,$$

which clearly exhibits shrinkage but no thresholding. The lasso penalized log-likelihood

$$-L(\boldsymbol{\beta},\boldsymbol{\sigma}^2)+\lambda\sum_{i=1}^{m}\sigma_i \tag{17}$$

subject to nonnegativity constraint $\sigma_i \geq 0$ achieves both ends. The update of σ_i is chosen among the positive roots of a quartic equation and the boundary 0, whichever yields a lower objective value. Next section illustrates variance component selection using lasso penalty on a real genetic dataset.

7. A Numerical Example

Quantitative trait loci (QTL) mapping aims to identify genes associated with a quantitative trait. Current sequencing technology measures millions of genetic markers in study subjects. Traditional single-marker tests suffer from low power due to the low frequency of many markers and the corrections needed for multiple hypothesis testing. Region-based association tests are a powerful alternative for analyzing next-generation sequencing data with abundant rare variants.

Suppose $\boldsymbol{y}$ is an $n \times 1$ vector of quantitative trait measurements on n people, $\boldsymbol{X}$ is an $n \times p$ predictor matrix (incorporating

Table 4. Top 10 genes selected by the lasso penalized variance component model (17) are tallied with their marginal p-values in an association study of 200 genes and the complex trait `height`.

Lasso Rank	Gene	Marginal p-value	# Variants
1	DOLPP1	2.35×10^{-6}	2
2	C9orf21	3.70×10^{-5}	4
3	PLS1	2.29×10^{-3}	5
4	ATP5D	6.80×10^{-7}	3
5	ADCY4	1.01×10^{-3}	11
6	SLC22A25	3.95×10^{-3}	14
7	RCSD1	9.04×10^{-4}	4
8	PCDH7	1.20×10^{-4}	7
9	AVIL	8.34×10^{-4}	11
10	AHR	1.14×10^{-3}	7

predictors such as sex, smoking history, and principal components for ethnic admixture), and $\boldsymbol{G}$ is an $n \times m$ genotype matrix of m genetic variants in a predefined region. The LMM assumes

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{G\gamma} + \boldsymbol{\epsilon}, \quad \boldsymbol{\gamma} \sim N(\boldsymbol{0}, \sigma_g^2 \boldsymbol{I}), \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}_n),$$

where $\boldsymbol{\beta}$ are fixed effects, $\boldsymbol{\gamma}$ are random genetic effects, and σ_g^2 and σ_e^2 are variance components for the genetic and environmental effects, respectively. Thus, the phenotype vector $\boldsymbol{Y}$ has covariance $\sigma_g^2 \boldsymbol{G}\boldsymbol{G}^{\mathrm{T}} + \sigma_e^2 \boldsymbol{I}_n$, where $\boldsymbol{G}\boldsymbol{G}^{\mathrm{T}}$ is the kernel matrix capturing the overall effect of the m variants. Current approaches test the null hypothesis $\sigma_g^2 = 0$ for each region separately and then adjust for multiple testing (Lee et al. 2014; Zhou et al. 2016). Instead of this marginal testing strategy, we consider the joint model

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X\beta} + s_1^{-1/2} \boldsymbol{G}_1 \boldsymbol{\gamma}_1 + \cdots + s_m^{-1/2} \boldsymbol{G}_m \boldsymbol{\gamma}_m + \boldsymbol{\epsilon}, \\ \boldsymbol{\gamma}_i &\sim N(\boldsymbol{0}, \sigma_i^2 \boldsymbol{I}), \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}_n) \end{aligned}$$

and select the variance components σ_i^2 via the penalization (17). Here, s_i is the number of variants in region i, and the weights $s_i^{-1/2}$ put all variance components on the same scale.

We illustrate this approach using the COPDGene exome sequencing study (Regan et al. 2010). After quality control, 399 individuals and 646,125 genetic variants remain for analysis. Genetic variants are grouped into 16,619 genes to expose those genes associated with the complex trait `height`. We include `age`, `sex`, and the top 3 principal components in the mean effects. Because the number of genes vastly exceeds the sample size $n = 399$, we first pare the 16,619 genes down to 200 genes according to their marginal likelihood ratio test p-values and then carry out penalized estimation of the 200 variance components in the joint model (17). This is similar to the sure independence screening strategy for selecting mean effects (Fan and Lv 2008). Genes are ranked according to the order they appear in the lasso solution path. Table 4 lists the top 10 genes together with their marginal LRT p-values. Figure 1 in the supplementary materials displays the corresponding segment of the lasso solution path. It is noteworthy that the ranking of genes by penalized estimation differs from the ranking according to marginal p-values. The same phenomenon occurs in selection of highly correlated mean predictors. This penalization approach for selecting variance components warrants further theoretical study.

8. Discussion

The current article leverages the MM principle to design powerful and versatile algorithms for variance components estimation. The MM algorithms derived are notable for their simplicity, generality, numerical efficiency, and theoretical guarantees. Both ordinary MLE and REML are apt to benefit. Other extensions are possible. In nonlinear models (Bates and Watts 1988; Lindstrom and Bates 1990), the mean response is a nonlinear function in the fixed effects $\boldsymbol{\beta}$. One can easily modify the MM algorithms to update $\boldsymbol{\beta}$ by a few rounds of Gauss–Newton iteration. The variance components updates remain unchanged.

One can also extend our MM algorithms to elliptically symmetric densities

$$f(\boldsymbol{y}) = \frac{e^{-\frac{1}{2}\kappa(\delta^2)}}{(2\pi)^{\frac{n}{2}} (\det \boldsymbol{\Omega})^{\frac{1}{2}}}$$

defined for $\boldsymbol{y} \in \mathbb{R}^n$, where $\delta^2 = (\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Omega}^{-1} (\boldsymbol{y} - \boldsymbol{\mu})$ denotes the Mahalanobis distance between $\boldsymbol{y}$ and $\boldsymbol{\mu}$. Here, we assume that the function $\kappa(s)$ is strictly increasing and strictly concave. Examples of elliptically symmetric densities include the multivariate t, slash, contaminated normal, power exponential, and stable families. Previous work (Huber and Ronchetti 2009; Lange and Sinsheimer 1993) has focused on using the MM principle to convert parameter estimation for these robust families into parameter estimation under the multivariate normal. One can chain the relevant majorization $\kappa(s) \le \kappa(s^{(t)}) + \kappa'(s^{(t)})(s - s^{(t)})$ with our previous minorizations and simultaneously split variance components and pass to the more benign setting of the multivariate normal. These extensions are currently under investigation.

Acknowledgments

The authors thank Michael Cho, Dandi Qiao, and Edwin Silverman for their assistance in processing and assessing COPDGene exome sequencing data.

Funding

The research is partially supported by NIH grants R01HG006139, R01GM53275, R01GM105785 and K01DK106116. COPDGene is supported by NIH grants R01HL089897 and R01HL089856.

References

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015), "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, 67, 1–48. [353]

Bates, D., and Pinheiro, J. (1998), "Computational Methods for Multilevel Models," Technical Report Technical Memorandum BL0112140-980226-01TM, Murray Hill, NJ: Bell Labs, Lucent Technologies. [350,358]

Bates, D. M., and Watts, D. G. (1988), *Nonlinear Regression Analysis and Its Applications*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: Wiley. [360]

Bien, J., and Tibshirani, R. J. (2011), "Sparse Estimation of a Covariance Matrix," *Biometrika*, 98, 807–820. [351]

Borg, I., and Groenen, P. J. (2005), *Modern Multidimensional Scaling: Theory and Applications*. New York: Springer Science & Business Media. [351]

Boyd, S., and Vandenberghe, L. (2004), *Convex Optimization*. Cambridge: Cambridge University Press. [351,352]

Callanan, T. P., and Harville, D. A. (1991), "Some New Algorithms for Computing Restricted Maximum Likelihood Estimates of Variance Components," *Journal of Statistical Computional Simulation*, 38, 239–259. [350]

De Leeuw, J. (1994), "Block-Relaxation Algorithms in Statistics," in *Information Systems and Data Analysis*, eds. H.-H. Bock, W. Lenski, and M. M. Richter, New York: Springer, pp. 308–324. [351]

Demidenko, E., and Massam, H. (1999), "On the Existence of the Maximum Likelihood Estimate in Variance Components Models." *Sankhyā*, Series A, 61, 431–443. [353]

Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum Likelihood from Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Soceity*, Series B, 39, 1–38. [350,351]

Ding, J., Tian, G.-L., and Yuen, K. C. (2015). "A New MM Algorithm for Constrained Estimation in the Proportional Hazards Model," *Computational Statistical & Data Analysis*, 84, 135–151. [351]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space" (with discussion), *Journal of the Royal Statistical Society*, Series B, 70, 849–911. [360]

Golub, G. H., and Van Loan, C. F. (1996), *Matrix Computations*, Johns Hopkins Studies in the Mathematical Sciences (3rd ed.), Baltimore, MD: Johns Hopkins University Press. [352]

Grzadziel, M. and Michalski, A. (2014), "A Note on the Existence of the Maximum Likelihood Estimate in Variance Components Models," *Iscussiones Mathematicae Probability and Statistics.*, 34, 159–167. [353]

Gupta, A., and Nagar, D. (1999), *Matrix Variate Distributions*, Monographs and Surveys in Pure and Applied Mathematics, Boca Raton, FL: Chapman & Hall/CRC Press. [356]

Hartley, H. O., and Rao, J. N. K. (1967), "Maximum-likelihood Estimation for the Mixed Analysis of Variance Model," *Biometrika*, 54, 93–108. [350]

Harville, D., and Callanan, T. (1990), "Computational Aspects of Likelihood-based Inference for Variance Components," in *Advances in Statistical Methods for Genetic Improvement of Livestock*, eds., D. Gianola and K. Hammond, Vol. 18 of *Advanced Series in Agricultural Sciences*. Berlin: Springer, pp. 136–176. [350]

Harville, D. A. (1977), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," *Journal of American Statistical. Association*, 72, 320–340. With a comment by J. N. K. Rao and a reply by the author. [350]

Heiser, W. J. (1995), "Convergent Computation by Iterative Majorization: Theory and Applications in Multidimensional Data Analysis," *Recent Advances in Descriptive Multivariate Analysis*, 157–189. [351]

Horn, R. A., and Johnson, C. R. (1985), *Matrix Analysis*. Cambridge: Cambridge University Press. [352]

Huber, P. J., and Ronchetti, E. M. (2009), *Robust Statistics*, Wiley Series in Probability and Statistics (2nd ed.), Hoboken, NJ: Wiley. [360]

Hunter, D. R. (2004), "MM Algorithms for Generalized Bradley–Terry Models," *The Annals of Statistics*, 32, 384–406. [351]

Hunter, D. R., and Lange, K. (2002), "Computing Estimates in the Proportional Odds Model," *Annals of the Institute of Statistical Mathematics*, 54, 155–168. [351]

——— (2004), "A Tutorial on MM Algorithms," *The American Statistician*, 58, 30–37. [351]

Hunter, D. R., and Li, R. (2005), "Variable Selection Using MM Algorithms," *Annals of Statistics*, 33, 1617–1642. [351]

Jeon, M. (2012), "Estimation of Complex Generalized Linear Mixed Models for Measurement and Growth," PhD thesis, University of California, Berkeley. [351]

Kiers, H. A. (2002), "Setting Up Alternating Least Squares and Iterative Majorization Algorithms for Solving Various Matrix Optimization Problems," *Computational Statistics & Data Analysis*, 41, 157–170. [351]

Laird, N., Lange, N., and Stram, D. (1987), "Maximum Likelihood Computations With Repeated Measures: Application of the EM Algorithm," *Journal of American Statistical Association*, 82, 97–105. [350]

Laird, N. M., and Ware, J. H. (1982), "Random-Effects Models for Longitudinal Data," *Biometrics*, 38, 963–974. [350,358]

——— (2010), *Numerical Analysis for Statisticians*, Statistics and Computing (2nd ed.), New York: Springer. [355,358]

Lange, K. (2016), *MM Optimization Algorithms*. Philadelphia, PA: Society for Industrial and Applied Mathematics. [351]

Lange, K., Hunter, D. R., and Yang, I. (2000), "Optimization Transfer Using Surrogate Objective Functions," *Journal of Computational and Graphical Statistics*, 9, 1–59. With discussion, and a rejoinder by Hunter and Lange. [351]

Lange, K., Papp, J., Sinsheimer, J., Sripracha, R., Zhou, H., and Sobel, E. (2013), "Mendel: The Swiss Army Knife of Genetic Analysis Programs," *Bioinformatics*, 29, 1568–1570. [353]

Lange, K., and Sinsheimer, J. S. (1993), "Normal/independent Distributions and Their Applications in Robust Regression," *Journal of Computational and Graphical Statistics*, 2, 175–198. [360]

Lange, K., and Zhou, H. (2014), "MM Algorithms for Geometric and Signomial Programming," *Mathematical Programming*, Series A, 143, 339–356. [351]

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014), "Rare-variant Association Analysis: Study Designs and Statistical Tests," *The American Journal of Human Genetics*, 95, 5–23. [360]

Lindstrom, M. J., and Bates, D. M. (1988), "Newton–Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data," *Journal of American Statistical Association*, 83, 1014–1022. [350]

——— (1990), "Nonlinear Mixed Effects Models for Repeated Measures Data," *Biometrics*, 46, 673–687. [350,360]

McLachlan, G. J. and Krishnan, T. (2008), *The EM Algorithm and Extensions*, Wiley Series in Probability and Statistics (2nd ed.), Hoboken, NJ: Wiley-Interscience. [358]

Meng, X.-L., and Rubin, D. B. (1991), "Using EM to Obtain Asymptotic Variance–Covariance Matrices: The SEM Algorithm," *Journal of the American Statistical Association*, 86, 899–909. [355]

Pinheiro, J. and Bates, D. (1996), "Unconstrained Parametrizations for Variance–Covariance Matrices," *Statistics and Computing*, 6, 289–296. [356]

Rao, C. R. (1973), *Linear Statistical Inference and its Applications* (2nd ed.), New York: Wiley. [352]

Regan, E. A., Hokanson, J. E., Murphy, J. R., Make, B., Lynch, D. A., Beaty, T. H., Curran-Everett, D., Silverman, E. K., and Crapo, J. D. (2010), "Genetic Epidemiology of COPD (COPDGene) Study Designs," *COPD*, 7, 32–43. [360]

Schafer, J. L., and Yucel, R. M. (2002), "Computational Strategies for Multivariate Linear Mixed-Effects Models with Missing Values," *Journal of Computational and Graphical Statistics*, 11, 437–457. [350]

Schelldorfer, J., Bühlmann, P., and van de Geer, S. (2011), "Estimation for High-dimensional Linear Mixed-Effects Models Using ℓ_1-Penalization," *Scandinavian Journal of Statistics*, 38, 197–214. [359]

Schur, J. (1911). "Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen," *Journal Fur Die Reine Und Angewandte Mathematik*, 140, 1–28. [352]

Wu, T. T., and Lange, K. (2010), "The MM Alternative to EM," *Statistical Science*, 25, 492–505. [351]

Yen, T.-J. (2011), "A Majorization–Minimization Approach to Variable Selection Using Spike and Slab Priors," *Annals of Statistics*, 39, 1748–1775. [351]

Yu, Y. (2010), "Monotonic Convergence of a General Algorithm for Computing Optimal Designs," *Annals of Statistics*, 38, 1593–1606. [351]

Zhou, H., and Lange, K. (2010), "MM Algorithms for Some Discrete Multivariate Distributions," *Journal of Computational and Graphical Statistics*, 19, 645–665. [351]

Zhou, J. J., Hu, T., Qiao, D., Cho, M. H., and Zhou, H. (2016), "Boosting Gene Mapping Power and Efficiency with Efficient Exact Variance Component Tests of Single Nucleotide Polymorphism Sets," *Genetics*, 204, 921–931. [360]